Scientia Iranica D (2013) 20 (3), 657–662

Sharif University of Technology

Scientia Iranica

Transactions D: Computer Science & Engineering and Electrical Engineering

www.sciencedirect.com

A combinatorial cooperative-tabu search feature reduction approach

E. Ansari [a,b,*], **M.H. Sadreddini** [a], **B. Sadeghi Bigham** [b], **F. Alimardani** [a,b]

[a] *School of Engineering, Computer Sciences and Engineering, Campus No.2, Shiraz University, Shiraz, 71348-51154, Iran*

[b] *Institute for Advanced Studies in Basic Sciences, GavaZang, Zanjan, 45137-66731, Iran*

Received 22 July 2011; revised 28 June 2012; accepted 16 October 2012

KEYWORDS

Tabu search;
Filter and wrapper;
Mutual information;
Nearest neighbor classifier;
Voronoi diagram.

Abstract Presenting an efficient general feature selection method for the problem of the curse of dimensionality is still an open problem in pattern recognition, and, considering the cooperation among features through search processes, it is the most important challenge. In this paper, a combinatorial approach has been proposed, which consists of three feature reduction algorithms that have been applied in a parallel manner to cooperate. We consider each of these algorithms as a component in a reduction framework. For each component, among all various attribute selection algorithms, the Tabu Search (TS) a useful and state of the art algorithm, is used. To take account of the interaction between features, more subsets should be examined. Hence, each component should explore individually through feature space in a local area which is different from other components. The proposed algorithm, called the Cooperative-Tabu-Search (CTS), and also a revised version of this new method, is introduced to accelerate the convergence. After sufficient iterations, which satisfy the objective function; the final subset has been selected by voting between three reduction phases, and the data is then transformed into the new space, where the data are classified with some commonly used classifiers, such as Nearest Neighbor (NN) and Support Vector Machine (SVM). The employed benchmark of this paper is chosen among the UCI datasets to evaluate the proposed method compared to others. The experimental results show the supremacy of the accuracy of the implemented combinatorial approach in comparison with traditional methods.

1. Introduction

The curse of dimensionality, introduced by Bellman, is one of the most important problems in data classification with large input dimensions. Feature selection and feature extraction are common solutions [1–3].

In this article, an investigative approach has been defined to use the advantages of feature selection methods. Here, the aim is to use an appropriate way to map the feature vectors to a new space with lower dimensions and then to classify the test data by Nearest Neighbor (NN) and SVM classifiers. Practically, it is observed that in high dimensional problems, some attributes have noisy-values, and this can inadvertently reduce the accuracy rate of classification by affecting the gradient of the mapping hyper plane (a plane in 2-D). Hence, a combination of three parallel component feature reduction algorithms is proposed here to soften these effects, as illustrated in Figure 1. In the first step, a different initial condition is assigned to each tabu search. Then, the outcome of each strategy is given to a voting function to select the best subset by considering the cooperation between them.

The Feature Subset Selection (FSS) requires two metrics: first, a search strategy to select candidate subsets and second, an objective function to evaluate these candidates and return their "goodness" value. There is also a feedback signal used by the search strategy to select new candidates. Generally, exponential, sequential and randomized algorithms are used as search strategies to select a subset of features [4] and, in sequential algorithms, the forward and backward selections are defined by [5], which both start with an initial subset and which sequentially add/remove the feature that locally optimizes the objective function. In contrast, exhaustive searches are costly and

* Corresponding author at: School of Engineering, Computer Sciences and Engineering, Campus No.2, Shiraz University, Shiraz, 71348-51154, Iran. Tel.: +98 916 6129978.
E-mail address: ansari@cse.shirazu.ac.ir (E. Ansari).

Peer review under responsibility of Sharif University of Technology.

Production and hosting by Elsevier

1026-3098 © 2013 Sharif University of Technology. Production and hosting by Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.scient.2012.12.020

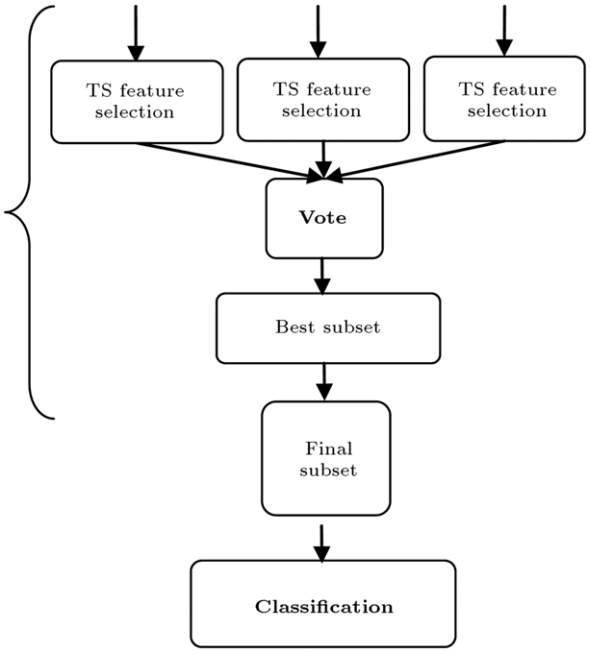

Figure 1: The three components feature reduction selects the best subset of features in each epoch and then the data are given to classifier.

time consuming, but can give a global optima. The tabu search that is represented by Ke, Jiang and De Ma [6], is an exhaustive search approach, whose time complexity depends on the initial condition and, in practice, usually stops. Therefore, in this study, three different initializations have been assigned, and to hasten the executive time, voting between the best has been used.

Objective functions are generally divided into two groups: Filters and Wrappers [7]. From an experimental research point of view, it is obvious that both groups have some advantages and disadvantages [8]. In the filter approach, the feature subset selection is performed independently of the classifier training phase. In this case, feature subset selection is considered a preprocessing step for induction. Although this is computationally more efficient, it ignores the fact that an optimal selection of features is dependent on the classifier model. While the wrapper approach is more complex than the filter, the interaction between feature subset and classifier is outstanding here. In this new method, two types of filter approach have been used, based on their generality and adaptivity to data properties. Among the vast varieties of search strategies, the tabu search is chosen for its feature selection, which is discussed in Section 2.

The classification phase of reduced data is described in Section 3. Results and a discussion of applying the proposed methods on UCI-data sets are reported in Section 4.

2. Feature selection phase

Among all strategies for selecting a subset of features, the Tabu Search (TS) is one of the most common algorithms to achieve the best possible informative and discriminative features, and its uses and versatility have been amply demonstrated by successful applications in various problems [6,9–11]. The basic concept of TS, as described by Glover [12,13], is a meta-heuristic superimposed on another heuristic. The main idea behind TS is very simple. A memory forces the search to explore the search space, such that entrapment in local minima is avoided. Unlike simple hill-climbing search techniques, but like simulated annealing, the tabu search often moves from a current solution to one which is worse, with the expectation that this move will eventually lead to an even better solution. The efficiency of tabu search has been proven in many optimization problems. The basic concepts of the tabu search algorithm are explained below:

(a) *Terms and definitions*:
- *Local move: The process of generating a feasible solution to the problem which is related to the current solution.*
- *Tabu list: A list of previous solutions.*
- *Tabu conditions: A set of rules which are used to derive, from the tabu list, regions of the search space from which any solutions are forbidden.*
- *Aspiration conditions: A set of rules which override the tabu conditions, to ensure that certain favorable local moves are accepted.*

(b) *Steps of algorithm*
N and M parameters are adjustable.
1. *Start with an initial solution.*
2. *If the current solution is better than the best solution so far, store it as the new best solution.*
3. *Add the current solution to the tabu list; remove the oldest item on the tabu list if it contains more than N items.*
4. *Apply the local move M times to generate M putative solutions.*
5. *Rank the putative solutions by fitness.*
6. *If the highest ranked putative solution is better than the current solution, jump to Step* 8.
7. *Eliminate those putative solutions which satisfy the tabu conditions, unless they also satisfy the aspiration conditions.*
8. *Select the highest ranked putative solution that was not eliminated as the new current solution, unless all putative solutions were eliminated.*
9. *If the termination criterion is not satisfied, repeat from Step* 2.

Some evaluation functions are needed to implement the search. A filter and also a wrapper objective function are employed to show that the proposed method improves the accuracy of the classifier anyway. These objective functions are: Mutual information criterion, and the Davies–Bouldin index. In this paper, the mutual information between a feature and class label is used with negative sign. So, both measures should be minimized. To achieve this goal, the evaluation function should examine the objective value on different subsets of features until it gets close to a certain threshold value.

2.1. Davies–Bouldin index (DB)

The Davies–Bouldin index, introduced by Davies and Bouldin [15], gives a measure of separation of clusters. In the Davies–Bouldin index, the inter-cluster distance is weighted by cluster spread or variance. This measure is defined here: Let $X = \{x_1, \ldots, x_N\}$ be the data set and $C = (C_1, \ldots, C_K)$ be partitioned into K clusters. Let $d(x_i, x_j)$ be the distance between x_i and x_j. Then, the Davies–Bouldin index is defined by Eq. (1) [4,14,15]:

$$DB(C) = \frac{1}{K}\sum_{i=1}^{K} \max\left\{\frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)}\right\}_{i\neq j}, \quad (1)$$

where $\Delta(C_i)$ is the intra-cluster distance and $\delta(C_i, C_j)$ is the inter-cluster distance. In this article, the Davies–Bouldin distance has been used to measure the factor of distance between trained data classes. The lower is a factor; to represent the data, the selected subset of features is more convenient. The time complexity of the Davies–Bouldin index is linear in the number of classified patterns.

2.2. Mutual information

In the design of self-organizing systems, the primary objective is to develop an algorithm that is able to learn an input–output relationship of interest on the basis of input pattern alone [16,17]. In this context, the notation of mutual information is of profound importance because of some highly desirable properties. Consider a stochastic system with input X and output Y. Both X and Y are permitted to take *discrete* values only denoted by x and y, respectively. The entropy $H(X)$ is a measure of prior uncertainty about X, as described by Shannon and Weaver [18] and Kwak and Choi [19]. *How can we measure the uncertainty about X after observing Y*? In order to answer this question, the entropy of X, with respect to a given Y, is defined as Eq. (2) by Kwak and Choi [19,20]:

$$H(X|Y) = H(X, Y) - H(Y). \tag{2}$$

The conditional entropy $H(X|Y)$ represents *the amount of uncertainty remaining about the system input after the system output, Y, has been observed*. $H(X, Y)$ is the *joint entropy* of X and Y, which is defined by Eq. (3):

$$H(X, Y) = -\sum\sum p(x, y) \log p(x, y). \tag{3}$$

In Eq. (3), $p(x, y)$ is the *joint probability mass function* of discrete random variables, X and Y. Since the entropy $H(X)$ represents our uncertainty about the input of the system *before* observing the system output, and the conditional entropy $H(X|Y)$ represents our uncertainty about the input *after* observing the output, the difference $H(X) - H(X|Y)$ must represent our uncertainty about the system input that is resolved by observing the system output. This quantity is called the *"mutual information"* between random variables X and Y, and is defined as follows:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= \sum\sum p(x, y) \log(p(x, y/p(x)p(y))). \end{aligned} \tag{4}$$

Entropy is a special case of mutual information:

$$H(X) = I(X; X). \tag{5}$$

The mutual information between two discrete variables, X and Y, has the following properties:

1-The mutual information between X and Y is symmetric, that is:

$$I(Y; X) = I(X; Y), \tag{6}$$

where the mutual information $I(Y; X)$ is a measure of the uncertainty about the system output, Y, that is resolved by observing the system input, X, and the mutual information, $I(X; Y)$, is a measure of the uncertainty about the system input that is resolved by observing the system output.

2-The mutual information, X and Y, is always nonnegative, that is represented in Eq. (7):

$$I(X; Y) \geq 0. \tag{7}$$

In effect, this property states that the information cannot be lost on the average, by observing the system output, Y. Moreover, the mutual information is zero if, and only if, the input and output of the system are statistically independent.

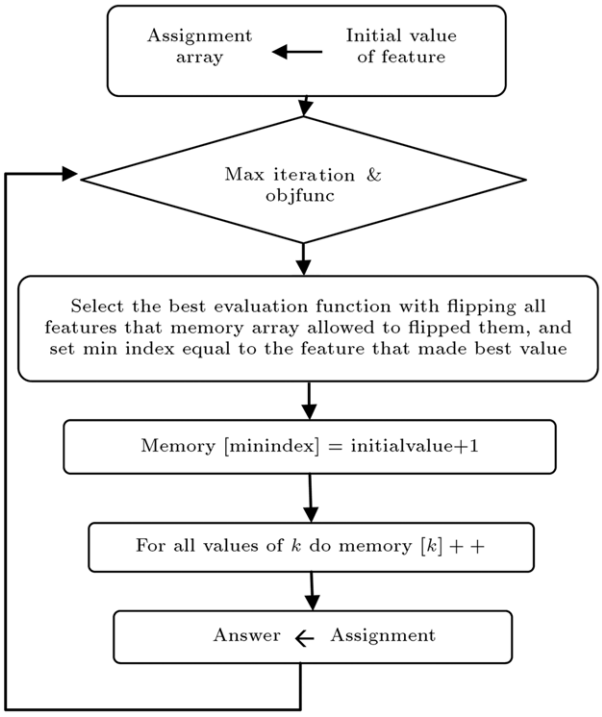

Figure 2: The flowchart of tabu search.

2.3. Presented "cooperative-tabu-search" algorithm

In this study, a new algorithm is introduced and entitled, the Cooperative-Tabu Search (CTS). In CTS, three TS components are initialized by a random subset of attributes. Then, each component uses one of the above aforementioned filter measures as an objective function to evaluate the goodness of each feature. Both these measures need a threshold value that should be tuned in the tabu search algorithm, and this best threshold value is found here using trial-and-error.

To implement the tabu search strategy, an array named assignment with size of input features is used and also a memory with an initial value, i.e. the number of tabu iterations. Through iterations, one bit of the assignment array flipped when the value of 1 means that this feature of the data is selected, and 0 means when it is not. Thus, one unit is decreased from the memory array. When a bit flips, it is not allowed to flip again in the next particular iteration, which gives an opportunity to explore more solutions. This particular forbidden iteration is arbitrarily chosen as five. This algorithm is summarized in the flowchart of Figure 2. The pseudo code of the feature selection is mentioned in Figure 3 in which the evaluate is a function that uses the DB or MI index to measure the goodness of the current subset (assignment).

2.4. Revised cooperative-tabu-search algorithm

As mentioned before, TS is an algorithm which can give the optimal solution if the threshold value for iteration is small. Besides, this algorithm is very time-consuming, and the epochs needed to converge to a certain subset highly depends on the initial subset. In the presented CTS algorithm, the voting between three components of TS hastens the convergence. The initial values in CTS are set randomly. This randomness would affect the convergence. If a statistical strategy can be used as the

```
Construct initial solution assignment (i) = 1 for all i
Construct initial memory, max try and best value = inf
for Iteration = 1: maxtry
    if (IsSatisfied(BestValue)) break
      for all features
          if (Memory(k) == 0|flag == 0)
             find j = Evaluate(Assignment)
          if (j <bestvalue)
               minval = j
                  minindex = nowindex
           if (IsSatisfied (j)) break;
      end for
      Memory (minindex) = initial + 1;
    for all features that Memory (i) is true
               Memory (i) = Memory (i) − 1;
      Not (Assignment(minindex));
      BestValue=min(minval,BestValue)
      Answer = Assignment
end for
```

Figure 3: Pseudo code of the feature selection.

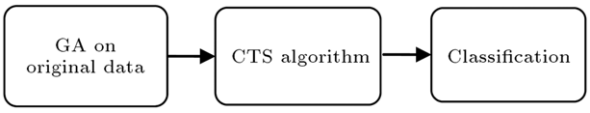

Figure 4: The three components feature reduction selects the best subset of features in each epoch and then the data are given to classifier.

initial phase, it can be more reliable. Among all strategies which consider the shape and distribution of data to find an initial good subset, the Genetic Algorithm (GA) is useful, as described by Zhang and Sun [21]. As illustrated in Figure 4, GA is used as a pre-processing phase to the CTS algorithm.

3. Classification

Common distance-based classification methods are naive bayes, decision tree, K-Nearest Neighbor (k-NN), regressive model, neural network and Support Vector Machine (SVM) [22], etc. Among them, NN and SVM are lately used in many different classification problems. Here, the reduced data are applied to these classifiers. What follows is a brief explanation of the mentioned classifiers.

3.1. Nearest neighbor classifier

The main idea behind the NN classifier method is to find a train pattern that it is the most similar pattern to the test. The nearest neighbour rule is a very intuitive method that classifies unlabeled examples based on their similarity to the examples in the training set. In this article, the Manhattan distance is used to compute dissimilarity. To get to know this distance, let X and Y be two patterns whose distances should be measured. Below, the Manhattan distance has been formulated:

The Manhattan or city-block distance is represented in Eq. (8), where D is the dimension of data:

$$\|X - Y\|_{c-b} = \sum_{k=1}^{D} |x_k - y_k|. \tag{8}$$

Table 1: Statistics of the data sets used in our computer simulations.

Dataset	Number of attribute	Number of patterns	Number of classes
WDBC	32	569	2
Wine	13	178	3
Glass	9	214	6
Sonar	60	208	2
Cancer	10	648	2
Image	19	210	7
Bupa	6	345	2
Yeast	8	1484	10
Australian	14	691	2
Satimage	36	6435	6
Vehicle	18	846	4

3.2. Support vector machine

The support Vector Machine (SVM) [23–25] implements The Structural Risk Minimization (SRM) principle. Underlying the success of SVM are mathematical foundations of statistical learning theory. Rather than simply minimizing the training error, SVM minimizes structural risk, which expresses an upper bound on a generalization error. Assuming a linear decision boundary, the central idea is to find a weight vector, W, such that the margin is as large as possible. Assuming that the data is linearly separable, an algorithm should seek to find the smallest possible W or maximum separation (margin) between the two classes. This can be formally expressed as a quadratic optimization problem, described in Eqs. (9) and (10):

$$\min_{w \neq 0, b} \frac{1}{2} \|w^2\|, \tag{9}$$

$$y_i(w^T x_i + b) \geq 1, \quad \forall i = 1, \ldots, n, \tag{10}$$

where $W^T X + b$ is the separation hyper plane between classes and X is input data. By transforming the above convex optimization problem into its dual problem, the solution can be found in the form of Eq. (11), where only a_i, corresponding to those data points which achieve equality constraints in Eq. (10), are non-zero. These data samples are called support vectors.

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{11}$$

SVM is a local method in the sense that the solution is exclusively determined by support vectors, whereas all other data points are irrelevant to the decision hyper plane.

4. Results

In this study, a new combinatorial algorithm is presented in order to select the most informative features in large datasets which are input to classification problems. The TS is a good estimation of an exhaustive search, because of which, has been chosen to be the basic feature selection approach. Since the initialization point and stopping criteria are critical parameters in TS that impact the final subsets, here, the results of three versions of TS are combined to have the smoothest estimation of the best subset. So, each reduction layer is initialized with a random subset. At the end of an iteration of TS, the best results of all three components are evaluated together and the best of them defines the best value till now. Choosing random initials gives the exploration property the CTS algorithm, and defining a desirable threshold for the objective function conducts the CTS to exploit to the nearly optimal subset. In a revised version

Table 2: Accuracy rate by applying NN on features which are selected by Revised-CTS using DB index, Across Group Variance (AGV), Forward & Backward Feature Selection (FFS and BFS) methods.

Data set	RCTS	AGV	FFS	BFS
WDBC	**95.00 ± 1.97**	91.07 ± 1.81	94.97 ± 1.75	94.73 ± 2.70
Wine	89.58 ± 2.98	74.28 ± 2.3	93.45 ± 2.40	**94.15 ± 2.4700**
Glass	**75.02 ± 2.02**	72.6 ± 2.21	70.38 ± 2.22	70.29 ± 1.22
Sonar	**92.18 ± 3.68**	83.96 ± 1.6200	87.21 ± 2.52	89.98 ± 1.02
Cancer	**89.01 ± 1.34**	74.92 ± 1.01	82.96 ± 1.41	83.44 ± 1. 12
Image	86.65 ± 4.5	83.34 ± 3.3902	84.36 ± 3.49	**87.96 ± 3.399**
Bupa	**61.9 ± 1.30**	56.04 ± 3.6710	54.15 ± 3.21	59.62 ± 1.35
Vehicle	63.8 ± 1.56	**66.4286 ± 3.1**	60.72 ± 1.78	63.64 ± 1.82
Satiamge	**91.47 ± 2.4**	89.14 ± 2.6	87.84 ± 1.41	88.21 ± 1.5
Yeast	**46.39 ± 1.46**	41.84 ± 2.3	42.11 ± 2.43	41.71 ± 2.57
Australian	**74.52 ± 1.5**	66.05 ± 0.8	63.34 ± 1.5	68.09 ± 0.312

Table 3: Accuracy rate by applying NN on features which are selected by Revised-CTS using MI index, Across Group Variance (AGV), Forward & Backward Feature Selection (FFS and BFS) methods.

Data set	RCTS	AGV	FFS	BFS
WDBC	94.50 ± 2.97	91.07 ± 1.81	94.37 ± 1.51	**94.73 ± 2.10**
Wine	90.51 ± 2.98	74.28 ± 2.3	93.45 ± 2.4900	**93.75 ± 2.75**
Glass	**78.90 ± 2.92**	74.16 ± 2.21	69.23 ± 1.20	72.09 ± 1. 5
Sonar	**92.18 ± 3.68**	83.96 ± 1.6200	87.21 ± 2.52	89.98 ± 2.502
Cancer	**87.21 ± 1.4**	73.20 ± 1.32	83.56 ± 1.01	82.40 ± 1.512
Image	83.15 ± 3.15	83.34 ± 3.3902	84.33 ± 3.49	**87.96 ± 3.399**
Bupa	**62.29 ± 1.00**	56.54 ± 2.71	52.55 ± 3.400	60.62 ± 1.35
Vehicle	**65.38 ± 1.56**	63.86 ± 3.1	60.22 ± 1.78	64.04 ± 2.82
Satiamge	**90.75 ± 2.4**	88.3514 ± 1.56	89.40 ± 1.1	86.11 ± 1.25
Yeast	44.9 ± 1.06	**45.7894 ± 2.3**	42.71 ± 2.43	42.11 ± 2.57
Australian	**72.20 ± 1.45**	69.75 ± 1.2	66.04 ± 2.6	68.09 ± 0.312

Table 4: Accuracy rate by applying SVM on features which are selected by Revised-CTS using DB index, Across Group Variance (AGV), Forward & Backward Feature Selection (FFS and BFS) methods.

Data set	RCTS	AGV	FFS	BFS
WDBC	**95.5 ± 1.72**	92.19 ± 0.81	93.37 ± 0.75	92.23 ± 1.2
Wine	89.08 ± 1.5	78.98 ± 0.53	92.45 ± 1.45	**93.85 ± 1.7**
Glass	**77.09 ± 1.9**	73.36 ± 1.51	70.53 ± 1.12	74.29 ± 1.22
Sonar	**90.68 ± 2.56**	85.17 ± 0.62	88.51 ± 1.52	89.88 ± 2.42
Cancer	**86.21 ± 2.34**	73.20 ± 2.2	86.16 ± 0.31	84.63 ± 1.22
Image	85.16 ± 1.5	84.34 ± 2.32	85.93 ± 1.40	**87.94 ± 1.30**
Bupa	**62.52 ± 2.12**	56.94 ± 3.6	59.55 ± 2.40	60.1 ± 1.35
Vehicle	64.18 ± 1.53	63.48 ± 0.31	**64.72 ± 1.78**	64.4 ± 0.81
Satiamge	**90.15 ± 0.34**	89.94 ± 1.56	89.14 ± 1.21	87.11 ± 2.25
Yeast	46.02 ± 2.46	**47.84 ± 3.3**	44.1 ± 3.43	40.91 ± 2.45
Australian	**76.52 ± 1.34**	67.75 ± 3.2	69.94 ± 1.8	69.19 ± 2.31

Table 5: Accuracy rate by applying SVM on features which are selected by Revised-CTS using MI index, Across Group Variance (AGV), Forward & Backward Feature Selection (FFS and BFS) methods.

Data set	RCTS	AGV	FFS	BFS
WDBC	**95.5 ± 0.97**	92.09 ± 0.8551	94.1 ± 0. 5	91.3 ± 1.23
Wine	88.23 ± 1.98	78.28 ± 0.51	**92.95 ± 1.90**	93.45 ± 1.4
Glass	**78.19 ± 0.9**	71.36 ± 1.4521	69.63 ± 1.11	71.09 ± 1.00
Sonar	**91.38 ± 1.56**	86.17 ± 1.06	87.01 ± 0.5762	89.58 ± 0.45102
Cancer	**89.21 ± 0.34**	71.20 ± 0.2201	82.16 ± 0.3120	82.30 ± 0.2512
Image	86.65 ± 0.5	83.34 ± 0.3902	83.33 ± 0.4110	**87.94 ± 0.3990**
Bupa	**62.32 ± 1.12**	55.04 ± 0.6710	58.55 ± 0.4020	58.61 ± 0.3505
Vehicle	**65.08 ± 0.56**	64.4286 ± 31	62.32 ± 0.78	63.4 ± 0.8
Satiamge	**91.25 ± 0.34**	89.3514 ± 0.56	87.34 ± 0.21	85.21 ± 0.25
Yeast	45.32 ± 0.46	**47.94 ± 1.3**	43.1 ± 1.3	42.21 ± 0.45
Australian	**75.52 ± 0.34**	66.97 ± 0.2	67.34 ± 1.0	64.9 ± 1.31

of CTS, the initials are set purposive to increase exploitation. But the intimate structure of TS guarantees that the exploration property is considered adequately. So, a revised CTS (RCTS) can give the most subsets by using a good measure of exploration and exploitation together. To show the performance of the proposed method, 11 standard datasets from the UCI ML repository are used for a number of features ranging from 6 to 60 (Table 1).

For any of datasets, all samples available are divided into 90% training set, among the samples, and 10% testing set, based on

10-fold cross validation, based on the method described by Lau and Wu [26]. A comparison of the NN classifier on the selected feature, by our method, to other state-of-the-art algorithms, is shown in Tables 2 and 3, and the results of the SVM are available in Tables 4 and 5. The results show the efficiency of our presented method.

5. Conclusion

In this paper, a novel combinatorial feature selection framework was proposed. The new algorithms, CTS and RCTS, were implemented and evaluated through various UCI datasets compared with related feature selection algorithms. The feature selection results were further verified by applying two different classifiers to the data. Our method demonstrated its efficiency and effectiveness by defining cooperation between three different search lines through feature space. The cooperation was done by voting. The supremacy of our approach was reported in the tables, especially on the Glass dataset.

We will study the effect of wrapper criteria on the proposed framework in the future and also test divergence based criteria or a combination of filter and wrapper.

Acknowledgment

The authors gratefully acknowledge the contributions of Dr. Reza Boostani to this work.

References

[1] Cios, K.J., Pedrycz, W. and Swiniarski, R.W., *Data Mining Methods for Knowledge Discovery*, Kluwer Academic Publishers (Chapter 9) (1998).
[2] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis*, 1st Edn., John Wiley-Sons, New York (1973).
[3] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, 2nd Edn., Academic Press (1990).
[4] Molina, L.C, Belanche, L. and Nebot, À. "Feature selection algorithms: a survey and experimental evaluation", *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM'02, pp. 306–313 (2002).
[5] Aha, D.W. and Banker, R.L. "A comparative evaluation of sequential feature selection algorithms", *Proceeding of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 1–7 (1995).
[6] Ke, Q., Jiang, T. and De Ma, S. "A tabu search method for geometric primitive extraction", *Pattern Recognition Letter*, 18(14), pp. 1443–1451 (1997).
[7] Kohavi, R. and John, G.H. "Wrappers for feature subset selection", *Artificial Intelligence*, 97, pp. 273–324 (1997).
[8] Zhu, M. and Martinez, A.M. "Subclass discriminant analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8), pp. 1274–1286 (2006).
[9] Tahir, M.A., Bouridane, A. and Kurugollu, F. "Simultaneous feature selection and feature weighting using hybrid tabu search/K-nearest neighbor classifier", *Pattern Recognition Letter*, 28(4), pp. 438–446 (2007).
[10] Pan, S.M. and Cheng, K.S. "An evolution-based tabu search approach to codebook design", *Pattern Recognition*, 40(2), pp. 476–491 (2007).
[11] Zhang, H. and Sun, G. "Feature selection using tabu search method", *Pattern Recognition*, 35(3), pp. 701–711 (2002).
[12] Glover, F. "Tabu search: A tutorial interfaces", In *Center for Applied Artificial Intelligence*, 20 (4), University of Colorado, Boulder, Colorado, pp. 74–94 (1990).
[13] Glover, F. and Laguna, M., *Tabu Search, in Modern Heuristic Techniques for Combinatorial Problems*, C.R. Reeves, Ed., John Wiley & Sons, Inc. (1993).
[14] Bolshakova, N. and Azuaje, F. "Cluster validation techniques for genome expression data", *Signal Processing*, 83, pp. 825–833 (2003).
[15] Davies, D. and Bouldin, D. "A cluster separation measure", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1(2), pp. 224–227 (1979).
[16] Battiti, R. "Using mutual information for selecting features in supervised neural net learning", *IEEE Transaction on Neural Networks*, 5(4), pp. 537–550 (1994).
[17] Sheng, Y. and Gu, J. "Feature selection based on mutual information and redundancy-synergy coefficient", *Institute of Image Processing & Pattern Recognition*, 5(11), pp. 1382–1391. Shanghai Jiaotong University (2004).
[18] Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL (1949).
[19] Kwak, N. and Choi, C.H. "Improved mutual information feature selector for neural networks in supervised learning", *International Joint Conference on Neural Networks*, IJCNN '99, 2, Washington, DC, pp. 1313–1318 (1999).
[20] Kwak, N. and Choi, C.H. "Input feature selection by mutual information based on parzen window", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(12), pp. 1667–1671 (2002).
[21] Zhang, H. and Sun, G. "Feature selection using tabu search method", *Pattern Recognition*, 35(3), pp. 701–711 (2002).
[22] Vapnic, V.N., *Statistical Learning Theory*, Wiley, New York (1998).
[23] Cherkassky, V. and Mulier, F., *Learning from Data: Concepts, Theory and Methods*, Wiley Interscience (1998).
[24] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer Verlag, Berlin (1995).
[25] Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D. and Vandewalle, J., *Least Squares Support Vector Machines*, World Scientific Publishing Co., Singapore (2002).
[26] Lau, K.W. and Wu, Q.H. "Leave one support vector out cross validation for fast estimation of generalization errors", *Pattern Recognition Letter*, 37(9), pp. 1835–1840 (2004).

Ebrahim Ansari received his B.S. degree in Computer Science from Yazd University, Iran, in 2005, and his M.S. degree in Computer Engineering from the Engineering School of Shiraz University, Iran, in 2009. He is currently a Ph.D. degree student in the Computer Science and Engineering Department at Shiraz University, Iran.

His research interests include: machine translation, association rule mining and distributed processing.

Mohammad Hadi Sadreddini is Associate Professor in the Computer Science and Engineering Department at Shiraz University, Iran. He received his B.S. degree in Computer Science in 1985, his M.S. degree in Information Technology in 1986 and a Ph.D. degree in Distributed Information Technology, in 1991, from Ulster University, in the UK. He has been working in Shiraz University, Iran, since 1993. His research interests include: association rule mining, bioinformatics, and machine translation.

Bahram Sadeghi Bigham obtained his B.S. degree in Mathematics from Birjand University, Iran. He completed his M.S. and Ph.D. degrees at Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2000 and 2008, respectively. He is currently Assistant Professor in Computer Sciences and Dean of the Department of Computer and Information Sciences at the Institute for Advanced Studies in Basic Sciences (IASBS), where he is recognized as the founder and director of the RoboScience Laboratory. Prior to arriving at IASBS, Dr. Sadeghi worked as a Postdoctoral Fellow at the University of Cardiff in the School of Computer Science. His research interests include: algorithms, computational geometry, data mining, artificial intelligent and e-learning.

Fatemeh Alimardani was born in Neyriz, Iran in 1983. She graduated in Computer Engineering, in 2009, from the Engineering School of Shiraz University (Artificial Intelligence), where she is currently a Ph.D. degree student. Her research interests include: signal processing, pattern recognition and statistical EEG analysing.